Full Name:______________________________________

CS 15-681, 15-781 Fall 1998

Final Exam

December 1998

Instructions:

- Make sure that your exam is not missing any sheets, then write your name *on every page indicated*.
- This exam is open book, open notes.
- You have three hours to take this exam.
- Write your answers in the space provided. If you need extra space, use the back of the preceding sheet.
- Write clearly and be concise.

Full Name:_______________________________________

Problem 1. (points): Miscellaneous

A. Suppose H is a set of possible hypotheses and D is a set of training data. We would like our program to output the most probable hypothesis h from H, given the data D. Under what conditions does the following hold?

$$\operatorname*{argmax}_{h \in H} P(h|D) = \operatorname*{argmax}_{h \in H} P(D|h)$$

B. Name 2 similarities and 2 differences between Boosting (you may assume AdaBoost) and Bagging.

C. Explain in your own words why learning from examples is futile without some form of inductive bias.

D. For each of the following algorithms, (a) state the objective that the learning algorithm is trying to optimize, and (b) indicate whether the algorithm is guaranteed to find the global optimum hypothesis with respect to this objective.

- Backpropagation with multi-layer networks
  - Objective:
  - Global optimum?:
- The perceptron training rule applied to a single perceptron
  - Objective:
  - Global optimum?:
- The FindS algorithm from Chapter 2
  - Objective:
  - Global optimum?:
- Support Vector Machines.
  - Objective:
  - Global optimum?:
- The EM algorithm
  - Objective:
  - Global optimum?:

E. In one sentence each, give

- one advantage of ID3 over Backpropagation
- one advantage of Backpropagation over ID3
- one advantage of FOIL over ID3
- one advantage of Support Vector Machines over Perceptrons

F. Notice in the PAC learning result given by equation 7.2, the number m of required examples grows without bound as we set ϵ closer to zero. Explain in plain English (no equations please) why the learner cannot learn a zero error hypothesis in the PAC learning setting, even though it can if it gets to pose queries to a trainer.

Problem 2. (points): Naive Bayes

A. Consider a learning problem where each instance x is described by the boolean attributes $a_1 \ldots a_n$, and where the target function $f : X \rightarrow V$ has two possible values: v_1 and v_2. Write the decision rule learned by a Naive Bayes learner *in terms of an inequality.* Circle the parameters of this decision rule that are estimated from the training data.

B. What is the form of the hypothesis space H considered by a naive Bayes classifier for this problem? Show that H corresponds to a set of linear decision surfaces in a space with dimension $2n$. (Hint: start by taking the log of the above inequality.)

C. Given that the naive Bayes algorithm learns a linear decision surface in the Euclidean space $\Re^{2n}$, does this imply that the naive Bayes algorithm and the perceptron learning algorithm will learn the same hypothesis (assuming here that the perceptron has the same $2n$ inputs)? If so, explain how to construct the corresponding perceptron. If not, explain why the decision surfaces learned by naive Bayes and the perceptron learning algorithm can differ.

D. Can you use the PAC results for consistent learners to give a bound on the sample complexity for the above naive Bayes learner? If so, give it and explain the conditions under which it is correct. If not, explain the difficulty.

E. The PAC bounds we discussed are statements of the form

with probability $1 - \delta$, the (true) error of h is less than ...

.

This question asks you to determine analogous bounds on errors in the parameter estimates of the naive Bayes learning algorithm. Let $\hat{P}(v_j)$ represent the learner's estimate for $P(v_j)$, and $\hat{P}(a_i|v_j)$ represent its estimate for $P(a_i|v_j)$. Assume the learner is provided m training examples, including m_j examples with target value v_j, and m_{ij} examples with both target value v_j and $a_i = 1$. Derive statements of the form

with probability $1 - \delta$, the error in the estimate $\hat{P}(v_j)$ is less than ...

.

with probability $1 - \delta$, the error in the estimate $\hat{P}(a_i|v_j)$ is less than ...

.

Full Name:______________________________________

Problem 3. (points): A Random Learner

Your friend asks your expert advice on her new learning algorithm, called LARCH (Learn A Random Consistent Hypothesis). The input to LARCH is any finite hypothesis space H and any consistent set of of training examples D. The LARCH algorithm is:

LARCH (training data D, hypothesis space H)

```
1. pick a hypothesis $h$ at random from $H$
2. if $h$ is consistent with all training examples in $D$, then output $h$ and terminate, else
3. go to step 1.
```

A. She first wants to know whether it is possible to bound the error of the hypothesis output by LARCH, based on the number of training examples provided in D. What do you tell her? If yes, give the bound and justify why it applies to LARCH. If not, then explain why the PAC bounds on sample complexity do not apply.

B. The other algorithm your friend has tried is LACH (Learn All Consistent Hypotheses). LACH outputs the set of *all* hypotheses from H that are consistent with the training data D, then classifies each new instance by a vote of these consistent hypotheses (weighted equally). She has two questions about LACH and LARCH

(a) Can you describe some condition under which one of these methods would be expected to produce a smaller true classification error than the other?

(b) Can you bound the true classification error of LARCH in terms of the true classification error of LACH? If so, state precisely the assumptions that must be satisfied for this bound to hold.

Full Name:______________________________________

Problem 4. (points): Miscellaneous

Answer each question in the space provided. Be concise and precise.

A. Amazing Web Technologies, Ltd. has hired you as a consultant for their latest genetic algorithm project: learning to distinguish which web home pages belong to Republicans versus Democrats. Due to the proprietary nature of their software product, they are unable to reveal to you the details of the hypothesis encoding used by the GA. However, they are willing to tell you that the GA encodes its hypothesis using a bitstring containing exactly 20 bits. They also reveal that this algorithm is allowed to run indefinitely until it outputs a hypothesis that classifies every training example correctly.

Their question to you is this: how many training examples of the boolean target function "Republican web pages" must they provide in order to assure that with 85% probability their GA will find a hypothesis whose true error is less than 15%? Please answer below.

They now run their GA and produce a hypothesis. When they test it on a set of 130 new instances they find that it commits 20 errors. What is the 90% confidence interval (two-sided) for the true error rate of this hypothesis? Give a one-sentence justification for your answer.

What is the 95% one-sided interval (i.e., what is the upper bound U such that $error_{\mathcal{D}}(h) \leq U$ with 95% confidence)? Give a one-sentence justification.

B. Instance-based learning methods such as nearest-neighbor simply store the training data when it is presented, and delay processing until a query instance is presented. For this reason they are called *lazy* learning methods, as opposed to *eager* methods such as ID3 and C4.5 that construct a general hypothesis at training time. Given a new query instance x, lazy methods have the advantage that they can construct a local approximation of the target function for the region of interest (i.e., the region near x).

Consider the eager Sequential Covering rule learning method, and the "Learn-one-Rule" algorithm summarized on pages 276–278 (Figure 10.1 and Tables 10.1 and 10.2). Suggest a lazy variant of this algorithm that delays learning until it is given a new query instance. Given the query instance x, this algorithm should use the training data to construct *a single rule that predicts the target value for* x. Describe your algorithm below, and show a trace analogous to the one in Figure 10.1. Use your trace to illustrate how your algorithm would operate when given the query instance

$$x = \langle Outlook = Sunny,\ Temperature = Hot,\ Humidity = High,\ Wind = Strong \rangle$$

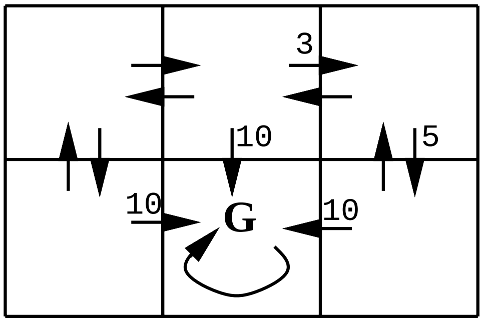
Does your lazy learning method ever classify instances differently than the eager algorithm described in the book? Explain your answer carefully.

Full Name:_______________________________________

Problem 5. (points): Reinforcement Learning

Consider the deterministic grid world shown below with the absorbing goal-state **G**. Here the immediate rewards $r(s, a)$ are 10, 5, or 3 for the labeled state-action transitions and 0 for all unlabeled transitions.

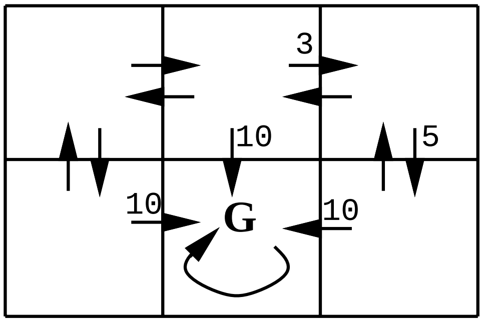
A. Write in the V^* values for each state in this grid world. Give the $Q(s, a)$ value for every transition. Finally, show an optimal policy. Use $\gamma = 0.8$.



B. Suggest a change to the reward function $r(s, a)$ that alters the $Q(s, a)$ values, but does not alter the optimal policy.

C. Suggest a change to $r(s, a)$ that alters $Q(s, a)$ but does not alter $V^*(s, a)$.

D. Now consider applying the Q learning algorithm to this grid world, assuming the table of $\hat{Q}$ values is initialized to zero. Assume the agent begins in the bottom left grid square and then travels clockwise around the perimeter of the grid until it reaches the absorbing goal state, completing the first training episode. Write in the new values of all $\hat{Q}$ values that are modified as a result of this episode. Answer the question again assuming the agent now performs a second identical episode (draw circles around the values you write in for this second episode, to distinguish them from the first values).



E. The task in reinforcement learning is to learn a policy $S \rightarrow A$ to choose an appropriate action a from the set A, given the current state s from the set S. The difficulty is that this is to be accomplished based only on indirect, delayed rewards. Q learning accomplishes this by instead learning an evaluation function over state-action pairs.

Suppose that you wish to learn the target function $S \rightarrow A$ directly, rather than learning an evaluation function. Consider Backpropagation, Genetic Algorithms, and Decision Tree learning. Which of these methods, if any, can be used to learn the function $S \rightarrow A$ from the kind of training experience that is provided to the Q learner in the above problem? For each of these three algorithms, explain why it is not possible, or sketch an approach.